DOCUMENT RESUME

ED 341 709 TM 017 888

AUTHOR Conley, Patrick; Jegerski, Jane
TITLE The Investigator Planning Exercise: The Selection of Detectives in the Chicago Police Department.
PUB DATE 91
NOTE 31p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Check Lists; Deduction; *Evaluators; Interrater Reliability; Investigations; Job Analysis; Job Applicants; Job Skills; Occupational Tests; Performance Tests; *Personnel Selection; *Police; Psychometrics; *Scoring; Screening Tests; Simulation; *Test Construction; Training; *Work Sample Tests
IDENTIFIERS *Chicago Police Department IL; *Investigator Planning Exercise

ABSTRACT

Construction of a work sample test, the Investigator Planning Exercise (IPE), for the job of detective in the Chicago (Illinois) Police Department is described. Simulated crime scenarios, a mock crime scene, and five checklists of necessary skills (i.e., ability to summarize and communicate facts, identify inconsistencies, and determine the next action) were prepared. To screen the maximum number of candidates, the IPE was designed to be administered to between 400 and 600 applicants in a single day by 50 boards of 3 raters each. Safeguards included using raters who did not know each other, assigning as many minority and female raters as possible to the boards, allowing the applicant to reject up to two boards for cause, and requiring raters to explain extremely divergent ratings on the six scales making up the evaluation. In all, 189 police sergeants received rater training, which began with a 5-hour classroom session examining a similar selection process. The process for youth officers was used. Rater trainees viewed videotapes of mock review boards and practiced rating applicants. Although raters were not given the actual checklists, they did receive explanations of restrictions on the rating process. Additional training was given in a 2-hour hands-on session in which the actual check lists were reviewed. Trainees were also offered optional training in the test process in 15 2-hour sessions introducing test-taking strategy. On the test day, 619 applicants were tested. The psychometric results are reviewed briefly. Four tables of study data and an eight-item list of references are included. (SLD)

THE INVESTIGATOR PLANNING EXERCISE:

THE SELECTION OF DETECTIVES IN THE CHICAGO POLICE DEPARTMENT

Patrick Conley
Chicago Police Department
University of Illinois at Chicago

&

Jane Jegerski
Elmhurst College

Address correspondence to:

Patrick Conley
T.J. O'Conner Training Center
1300 W. Jackson Blvd.
Chicago, IL 60607

Investigator Planning Exercise (IPE)

This paper describes the construction of a work sample test for the job of Detective in the Chicago Police Department. It has a number of different sections. It begins with a discussion of the need for a procedure that measured different types of job-related abilities identified in a job analysis phase.

Next, a description of the construction phase of the IPE is presented. This includes a description of the means used to generate a simulation that mirrored important aspects of virtually all actual police investigations conducted by both uniformed and detective personnel. Also included is a description of the verification phase that led to the construction of assessor rating forms as well as the method of setting scale weights.

Following this, there is a section that describes the method by which raters were selected and trained as well as the manner in which potential test contamination was averted.

This is followed by a discussion of the steps to insure applicant understanding of the nature and type of testing procedure.

Finally, the results of the IPE are examined in terms of psychometric characteristics. Here, we focus on the reliability and distributional properties of the procedure.

Need for an oral board exercise

During the evaluation of the job analysis data and the

construction of the job knowledge test, we noted that a number of important KSAs were not amenable to the multiple choice format. For example, such KSAs as the "Ability to identify consistencies in accounts of victims, etc." and "Ability to logically summarize the facts of an investigation" would be difficult to assess given a written format. Given our desire to test applicants' abilities in these areas, we found it necessary to examine alternative test methods.

After reviewing several formats, we decided to develop a structured work-sample exercise. This exercise would have panels of raters assess an applicant's ability to use the actual processes used by officers involved in criminal investigations.

There were a number of advantages to a structured work sample test. The first was that such a test could be designed in a way as to test a number of candidates in a relatively short time. (As will be seen later, 619 candidates were tested in a single day.)

Second, the method had shown its value in a previous youth officer selection procedure. These benefits included the acceptance by the department decision makers, those who actually completed the procedure, as well as those who were raters in the exercise (Conley, 1987).

Third, the youth procedure which used anchored checklists had the effect of increasing the amount of rater agreement (i.e., interrater reliability). For example, in this procedure there were only three times in the evaluation of 305 applicants where

there was the need for a raters' post-rating session to resolve disagreements in an applicant's final score. Further evidence was found when the average correlation between the three raters, over all boards, for all candidates on the various scales was .94 with a range of .91 to .98.

Fourth, an examination of the youth officer data provided evidence that the oral assessment exercise measured dimensions independent from those measured previously either in the written multiple choice test or in their performance grades. That is, we were measuring dimensions that were relatively independent.

Finally, the construction of a work sample exercise was based on the specific skills identified by supervisors as being important in the jobs of both detective and patrol officers.[1] This judgment was necessary because both the Uniform Guidelines on Employee Selection Procedures (EEOC, 1978) and the Principles for the Validation and Use of Personnel Selection Procedures (Division 14, APA; 1987) require that any content valid test construction strategy focus on skills that are presently possessed by an applicant and not those skills learned later in training or on the job.

Construction of the IPE

IPE Test Content

The 5 important KSAs were identified as meeting these two

---

[1] We obtained patrol officer supervisors' judgments of whether the KSAs identified as important by detectives were also judged as important for field officers.

criteria. These were:

1. The ability to logically summarize the facts of an investigation so that the facts can be presented to supervisors or Felony Unit personnel.
2. The ability to identify both consistencies and inconsistencies between the accounts of incidents given by victims, witnesses or offenders.
3. The ability to identify consistencies and inconsistencies between physical evidence and the accounts given by victims, witnesses, and offenders.
4. The ability to determine the direction of an investigation (i.e., which, of a number of courses of action, should be taken).
5. The ability to remember and communicate the facts of an investigation.

IPE Format Construction

Using the above KSAs as guidelines, the first author along with 6 first-line department supervisors began the construction of crime scenarios that would provide the stimuli for evaluation. The notion here was that actual scenarios would increase both the face validity and the sampling of representative behaviors needed in investigations.

The scenario construction which consisted of a number of steps started. Initially, we attempted to generate two different scenarios. Each was modeled after a particular type of crime (i.e., a home-invasion/rape and a burglary) and had certain characteristics.[2] First, each case contained a number of different reports (e.g., original case, supplementary reports, etc.). Each of these, in turn, had a number of discrepancies between statements of witnesses and that of the victim. Each

---

[2] It should be noted that while we initially attempted to build two scenarios later analysis indicated that the two were psychometrically different. The result of this was the construction of only the home-invasion/rape scenario.

also had a number of inconsistencies between physical evidence found on the scene and the accounts of victim and witnesses. Finally, the scenarios were written so that there were a number of investigative options available.

As construction efforts continued, it was necessary to evaluate the quality of the scenario as well as generate the answers that would be the basis of the scales mentioned earlier. To do this, several groups of detectives were brought together. Each group was given the scenario and told to do several things. First, they were instructed to read the scenario carefully taking notes that focused on the important aspects of the case. Next, they were asked to outline any discrepancies that they found between the statements of the witnesses and the victim. Finally, they were asked to list the things that they would do next in this investigation. This feedback information was gathered at the end of each session.

We next corrected and expanded the scenario so that it would mirror actual situations found on the job. This was done by first rewriting and making the editorial changes recommended by the first group of detective evaluators. This iterative process continued during the entire test building phase.

During this time, we located a mock crime scene. This was an apartment where counterfeit evidence was placed in a manner consistent with the plan of the scenario. For example, "evidence" consisting of "blood" was placed in different locations in the apartment which either agreed or disagreed with

work. Police officers who find agreement between various sources (i.e., physical evidence, statements of the witnesses, victims and offenders) must logically begin their search for an offender. It is only in those cases where there are inconsistencies between one or more of these sources of information is there a need to clarify the situation by going over the facts of the case.

Evidence supporting this notion was found when virtually all supervisors who participated in the construction phase noted that the identification of discrepancies was the first step in solving a crime. They noted that investigators are often cued to reinvestigate particular aspects of the crime because a victim's (or witness's) statement does not match either the evidence found at the scene or the accounts of others present.

Checklist 3 The third checklist would measure an applicant's ability to determine the course of an investigation. Here, the focus was on measuring an applicant's ability to pick one or more for continuing the investigation. Again, the object of this checklist was based on the notion that time and resources should not be wasted on either collecting or reexamining unimportant information.

Checklist 4 Finally, a fourth component was designed to measure an applicant's ability to communicate verbally. Specifically, this scale would insure that an applicant has some minimum capability of getting his/her point across.

It should be noted that Checklist 4 was to be used after each of the substantive scales. The reason for this was straight

forward. It was possible that due to the different types of information sought in each of the three above scales the nature of an applicant's response could be different. For example, an officer may be able to identify the important aspects of the scenario but less able to communicate his/her answers concerning a plan of action.

Checklist Weighting

The next step was to insure that the various checklists and their component items received the appropriate weighting. That is, it was necessary to weight each of the three checklists as well as the items within the checklist in such a way as to insure that the number of points achieved from each was commensurate with its relative importance.

To accomplish this, it was first necessary to match the relative weight for the particular checklist with a composite index representing the weight given to a KSA by supervisors during their rating sessions. This is essential because any testing procedure should be weighted in a manner that attempts to closely mirror the job analysis data (See Standards and Principles).

To accomplish this, we defined the KSA Index as the simple product of the KSA's value on two scales used in the KSA rating component of the job analysis. Specifically, KSA ratings of the percentage of supervisors who stated that the KSA was needed by new workers was multiplied by a scale rating of the amount of trouble new workers would have if they did not possess the KSA.

The rationale here was to give more weight to KSAs that are perceived as being more critical upon entry. The KSA Index was then summed across all the 5 important KSAs. The percentages associated with each of the component KSAs was computed by dividing the KSA Index for each by the sum of the 5 important KSAs. These percentages then became the checklist weightings for each scale. (See Table 1).

For example, Scale 1 (Presentation Content) has a total possible score of 29 which is 22.5% of all the points possible in the exercise. This is relatively close to the composite index for this KSA 1 (19.6). Similarly, Scale 3 (Inconsistencies in Witnesses' and Victim's Accounts) reflects that the percentage of the total possible score (37.6) closely matches the percentage of the KSA's index total for KSAs 2 & 3 (37.6). Scale 5 (Things to do) has an index of 240.35 which is 19.8 of the sum total of the

Table 1
Summary of KSA and Checklist weightings

KSA # KSA Description	KSA Index	% of KSA Total	Check list #	Total Scale Points Possible	% of Total Pts.
1. Ability to summarize facts	227.24	19.6	1	29	22.5
2 & 3 Ability to identify inconsistancies	434.84	37.6	3	50	38.8
4. Ability to determine next action	240.35	20.8	5	29	22.5
5. Ability to communicate facts	310.28	21.7	2,4,6	-21	16.2

index over KSAs. The percentage of total number of points available for Scale 5 versus the total number of points (22.5) closely matches the percentage of total points available (20.8). Finally, the percentages associated with communication skills generally reflect the total number of points associated with that those scales.

It was also necessary to set the weights of the various items within the various checklists. This was done by having another panel of incumbent detectives review the scenario and rate each item on each of the three checklists on a 5 point scale which ranged from "1" (not very important) to "5" (extremely important). For each response on each checklist, an average was computed. These averages of items within each checklist were then rank ordered according to their magnitude. Higher average items were given greater weight on the particular checklist.

In summary, the result of the above was that there were estimations of both the relative weight that should be given each checklist and the relative importance of each of the items within each checklists. This allowed for the weighting of both the items within a checklist and the checklists themselves.

Implementing the Investigator Planning Exercise (IPE)

The overall plan was to set a cut score on a written job knowledge test that would allow us to evaluate from 400 to 600 applicants. There were three reasons for this number. First, we needed enough applicants to fill our forecasted shortages. Second, there was a concern for having too many candidates. This

could lead to having an eligibility list that would last for years. Finally, our logistical ceiling limited us to a single testing site with a maximum of 50 boards.

In addition to these logistical constraints, there were also a number issues that needed specification. The included the need for test security, the identification and training of raters, and the ensuring of the standardization of the procedure for both raters and applicants.

Test Security

After analyzing various plans associated with other oral boards held by the city, it was decided that the IPE should be held on a single day. This plan would preclude the criticism that some applicants would be advantaged because they would have spoken to individuals who had already taken the test.

Applicants would be divided into two groups of no more than 310. Individuals in each test group would be to appear at either 8:00 AM or 11:30 AM. Applicants were assigned to one of the two times using an alphabetical list - A to L at 8:00 and L to Z at 11:30 AM. Those who were assigned to the early call would not be allowed to leave the test site until all those in the later group had been seated and locked in a general briefing area.

Rater Identification and Assignment

Because of the plan to have all applicants assessed in one day, we judged it necessary to have at least 50 boards each of

which would consist of 3 members.[3] Because of the large number of boards and the large number of applicants, it was necessary to build a procedure that minimized the opportunity of rater impropriety. The use of three member boards rested on the premise that agreement between the three members of a panel to illicitly inflate or deflate a particular applicant's score was dramatically more difficult than a situation where the board consisted of only two members.

Our attempt to dissuade raters from altering ratings was expected to be especially effective when certain safeguards were built into the system. First, there was a conscientious attempt to assign raters to a board who did not know one another. Rater groups were identified by unit and geographical location within the city. For most boards, raters were assigned on the basis of their units with raters coming from opposite sides of the city.

The second safeguard focused on the maximum use of minority raters. Specifically, we insured that minority and female raters were placed on as many boards as possible. The hope here was that female and minority applicants would have the opportunity to have a board which contained individuals from their own group. On the day of the test, applicants from these groups were asked if they desired boards that had minority or female raters. If they did they were sent randomly to one of the boards that had a

---

[3] Our original plan was to videotape each candidates performance without the use of a board. This would have allowed us to evaluate an applicant's performance using any number of raters. However, the costs of this plan was prohibitive.

similar group member. If they chose not to go before this type of board, they were randomly assigned to any board available.

The third safeguard concerned the applicant's ability to exclude up to two boards for cause. The rationale here was that there was some probability that a applicant might be assigned to a board that, for one reason or another, had one or more board members who were unsatisfactory.

This exclusion process was done by showing a list of sergeants that made up the assigned board. If an officer decided that one or more of the sergeants were unacceptable they had the option of being randomly assigned to a second board. Again, the applicant had the opportunity to review the raters on the second board. If there was some reason the applicant wanted another yet another board, a third board was randomly selected. Applicants had to take this third board.

Fourth, raters were told that ratee demographic information would not be collected until the end of the session, after all ratings had been made. Given the fact that ratees were required to wear civilian clothes, the hope here was that raters would not be able to cue on a particular name and distort ratings.

Finally, the procedure design forced raters to explain extremely divergent ratings on the six scales making up the evaluation. If raters' scale ratings were more that three points apart, a consensus meeting was declared necessary by the board chair. This, in turn, required that the two discrepant examiners inspect their individual item ratings within a scale and come to

an agreement. Checking the results of IPE, there were no cases where raters did not come to agree within this 3 point criterion.

Raters

Supervisors who were to do the IPE evaluations came from one of three groups. The first group consisted of all newly promoted sergeants who were previously assigned as detectives. The second were sergeants assigned to the Detective Division. Finally, any trained raters who have been involved with any of the city's previous assessment procedures were used. In total, 205 police sergeants were identified.

Training Sessions

Rater training associated with this exercise consisted of two phases. The first was a five hour classroom presentation that closely examined a procedure that was the immediate predecessor of the IPE - the youth officer selection procedure. This session examined the nature of work sample exercises in terms of its construction, use, and results expected. The second was a 2 hour training session on test day. This was designed to give raters the opportunity of a hands-on examination of the IPE.

Session 1

There were 10 five-hour training sessions. Each session trained 10-30 raters. The total of 189 supervisors were trained.

After a brief overview of the day's schedule, the training began with a description of both the scenario and the method by which it was developed. Next, a detailed description of what will be expected from both the candidate and the rater was

presented. To insure that supervisors knew what to expect on test day, raters were given both applicant and rater instructions from the new IPE as well as the stimulus material (i.e., the investigation, answer sheet and instructions for raters) from the previous youth officer exercise.

The decision to use the youth materials rested on two assumptions. First, there was a fear that training on the IPE would unnecessarily compromise the procedure. This was especially true given that training would be conducted as much as a month before the actual test day. Given this time, it was extremely likely that some supervisors would leak some facets of the procedure. Second, as noted earlier, the youth procedure was a first generation IPE; it was the conceptual basis for the IPE. As such, it provided a realistic tool that mirror the same sorts of rater tasks measured in the IPE.

At the training session, the supervisors were instructed to read and follow the IPE instructions that would be given to applicants and to board members on test day. Here, an attempt was made to obtain any feedback on any difficulty raters found in either set of instructions. All supervisors agreed that the instructions were both clear and comprehensive. (A copy of all forms and yet to be discussed test materials are available from the first author upon request.)

Next, the raters were given approximately 1 hour to examine the various documents associated with the youth officer procedure. They were asked to read the stimulus investigation

and examine the answer sheet. They were asked to critically evaluate the procedure and to question any area that they felt was problematic. This was followed by a discussion of any problem areas found. All problems were addressed and eventually resolved.

The supervisor raters were then asked to view a series of 7 videotaped simulations where current detective role players depicted candidates going before a mock oral board. The actors responded to the questions of the mock board members who followed the format of the IPE. Supervisors were asked to follow the responses of each applicant and actually rate individuals on their answer sheets. At the conclusion of each candidate's presentation, the raters' responses were checked. In virtually all rater classes, there was near unanimous agreement in scale scores between the rater trainees on an applicant's responses.

The value of these practice board ratings rested on the two perceptions of raters. The hope was to show raters that the use of a structured checklists would lead to: (1) more accurate ratings from the individual raters and (2) agreement between raters on the scale scores of an applicant. We apparently achieved this objective. Virtually all raters when asked by questionnaire stated that this rating method was extremely effective in terms of rater agreement and accuracy.

Rules associated with rater questions

Next, there was a discussion of a strict set of rules that limit the type of questions that raters could can ask of

applicants. Specifically, for sake of standardization, it was necessary to not only identify those situations where raters need to instruct applicants but also define the exact wording of these remarks.

The Checklists

Raters were also instructed that while they would not see the actual IPE checklists until test day there were a number of restrictions as to what they could do during the procedure. These written instructions focused on what the raters could say and how they were to use the particular lists. Raters were told to make their ratings independently and while the individual candidate is giving his/her answers.

Summary Checklist Ratings. At the end of an applicants presentation, raters were told that they would enter their ratings for each of the checklists on a summary sheet. The board chair would then collect the rating from each rater on each individual scale. These would be entered on a "Ratee Consensus Sheet" on the rear of the packets.

Consensus Ratings. Finally, raters were told of the mechanism that would allow for disagreements among raters to be resolved. This procedure known as a "consensus rating session" would be held when a scale rating from different raters did not agree within 3 points (i.e., the average rating would be larger than 1 scale point).

The consensus meeting would require the divergent raters to examine the applicant's responses on the particular checklist.

Raters were told that this discussion should continue for no longer than 10 minutes after which a simple average would be computed for the three separate ratings. (As will be seen later in the results section, there were no consensus meetings needed to resolve the rater differences.

Summary. After the examination of the checklists, raters were given a summary of what was expected of them. They were told of all the various logistic aspects associated with the test including such things as rater assignment, board makeup, and ratee assignment. The need for test security was again explained. Finally, raters were informed of the overall schedule for test day. Included here was a description of a two hour training session that would occur on test day.

Session 2

The second rater training session was held on test day and consisted of a two hour examination of the entire IPE. Each rater was given the scenario, the answer sheets, a packet of photos, both ratee and rater instructions and rater checklists. Raters were given time to read each of the IPEs components. Each handout was discussed. Problems were addressed and discussed.

Raters were then told of the method by which they could give bonus points to particular ratees for unusually good responses not listed in the scales. As will be seen later, this policy was reversed and with one exception all bonus points were removed.

At the end of the training session, raters were assigned to their respective boards and sent to their particular rooms.

IPE Ratee Familiarization

In addition to rater training, we also offered training to applicants who successfully passed the job knowledge test. There were two different training sessions. The first was a series of optional 2 hour sessions held about 2 weeks before test day. The second was a mandatory meeting held on test day immediately prior to the exercise.

Optional Ratee Training - Phase 1

In total, there were 15 different sessions scheduled. During these meetings, a number of issues were discussed. These included a detailed description of the test format and how this was different from the City's latest oral exercises. Timetables were discussed as were the locations of both study and test rooms.

Examinees were also told of how they would proceed from the initial orientation on test day through their individual study sessions. For example, ratees were told that after leaving their study rooms that they would go to a holding area. At this area, they were told that they would be able to exclude one or more raters from their particular board.

Ratees were informed of what the IPE was designed to measure. Specifically, the KSAs that were the basis of the IPE were discussed. They were informed that the IPE focused on three types of information. First, the IPE was designed to assess their ability to read and understand reports written by others. From this basis, ratees were expected to determine the important

aspects of an investigation. That is, they would be asked by the panel of supervisors to list the important aspects of the investigation. Second, applicants were also told that the IPE would focus on their ability to identify inconsistencies both in the accounts of various individuals and between the physical evidence found and these same accounts. Thirdly, applicants were also told that the IPE would require the identification of the courses of action they would do next.

It should be noted here that ratees were told that courses of action were limited to things that they would do immediately after receiving the mock investigation. For instance, arresting an offender would not be done. There were a number of preliminary steps that needed completion before this could be done. Similarly, ratees were told that procedural issues would not be an option on the test. For example, applicants were instructed to avoid any courses of action that could be learned after promotion. Such behaviors as completing reports, asking for complex tests for evidence, and completing court documents would not be on the raters answer sheets.

Applicants were told of a test taking strategy that could be used on test day. Specifically, they were told of a series of steps that would aid them in negotiating the IPE procedure. Finally, applicants were told that on test day, after they had reported, there would be another meeting where everything discussed in this first session would be gone over a second time.

Mandatory Ratee Training

On test day, two mandatory training sessions were held for applicants who were assigned to report at either 8:00 or 11:00 AM. Examinees at each of these sessions were given a sealed packet that contained their scenario, photos, plat, paper and pencil. They were also given a formal instruction packet which was explained in detail. Applicants were also told of the manner in which they would complete the entire procedure. Before leaving the assembly hall, applicants were given the opportunity to asks questions regarding any part of the procedure. All problems were resolved.

Testing Results

Check of IPE checklists

In the week that followed, the IPE scale ratings were checked for several different types of errors. First, all consensus sheets were checked to insure that the ratings met with the three point consensus rule. Individual scales were checked to insure the number of points given by each of the raters were properly recorded on the summary sheets. We found no problems.

Bonus Points. One major problem concerning the awarding of bonus points was found. As noted earlier, the IPE components, by design, allowed raters to given extra points for different responses not included in specific scales. For example, examining Scale 5, raters stated that completing a "name check of the victims previous criminal history" was a course of action that should be awarded 1 extra point.

The rationale for this extra point policy was based on the notion that raters would be capable of identifying different facets of the investigation that were omitted in the initial scale construction sessions. To insure that raters did not favor any one individual, it was necessary that raters awarding bonus points needed to have the other two panel members agree with the award of the points. This would be done by having any rater who believed that bonus points should be awarded enter the bonus behavior onto the space provided at the end of the scale. At the end of the rating session, the rater who wanted to give the bonus points needed only to bring up the response to his/her fellow board members. If all agreed that the response was worthy of extra points, they would decide on the number of points to be awarded and add these to the applicant's score on the particular scale.

Unfortunately after an examination of a number of test packets, we decided to eliminate bonus points. With one exception (i.e., name checking the victim in Scale 5 - given one bonus point), all bonus points were taken away. This was done because there were three types of evidence that indicated that the boards' were inconsistent in awarding points. First, some raters within the same board gave different numbers of points for the same courses of action given by different ratees. Though the number of occurrences was extremely small, this situation indicated that some ratees would be disadvantaged. Second, between board differences were found. Different boards gave

different amounts of points for the same course of action. Again, this difference was extremely small but nonetheless indicated a potential bias against some ratees. Finally, there was an apparent fatigue factor. Individuals who came at the early end of test day had a greater probability of receiving extra points than ratees who evaluated near the end of the session. (It should be noted, however, that there was no apparent fatigue factor in using the various scales. Rater agreement was constant over the entire time of rating.)

After all IPE protocols were checked and bonus points eliminated where necessary, the average of the three raters' scale values for each IPE scale was computed. The average scale values for all six scales were summed into a ratee's final score.

Psychometric Results

The final score distribution was then analyzed. Table 2 contains the descriptive statistics and plots (Tukey, 1977) for the various checklists of the IPE. Please note that the statistics for the communication checklists (i.e., scales 2, 4, and 6) are not listed. Our reason for this is that all but 3 individuals were not penalized for lack of communicative skills.

Table 3 contains the matrix of correlations between the various components of the IPE. As can be seen, the inter-scale correlations are small lending evidence that the dimensions are relatively independent.

Table 2
Descriptive statistics and plots for the components of the IPE

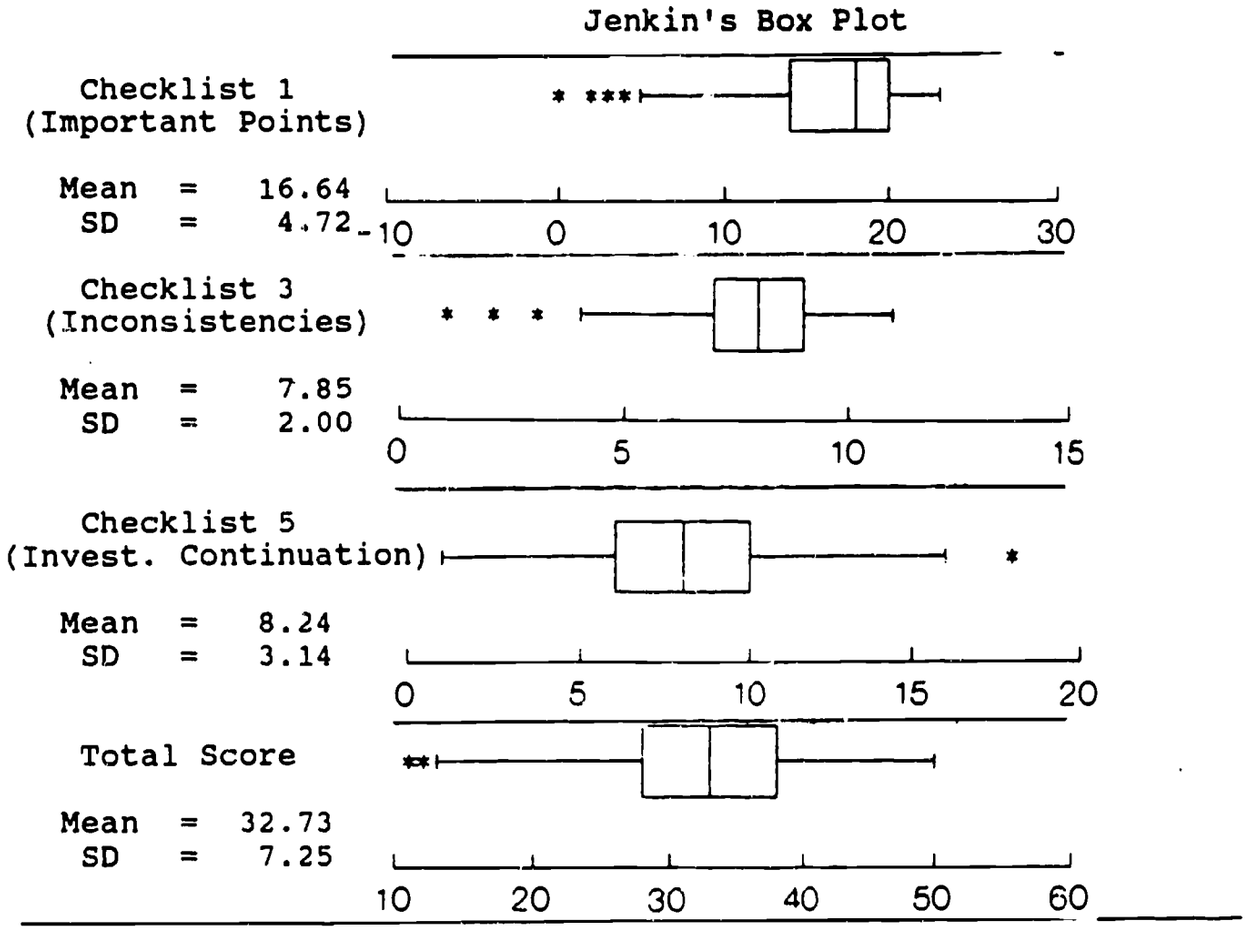

Table 3

Correlation matrix of the various IPE components

Scale	#			
Important points	1	1.00		
Discrepancies	3	.24	1.00	
Courses of action	5	.30	.24	1.00
		1	3	5

Table 4 in turn examines the internal reliability estimates for the entire test and for each component. The internal reliability for the total IPE is respectable regardless of the index used (i.e., .76, .87, and .83) (Allen & Yenn, 1979). The reliabilities for the different components are smaller. However, this pattern can be expected given the smaller number of items making up each of the particular scales. For example, the lowest reliabilities are those associated with Scale 2 (11 Items). They increase for Scale 3 (16 items) and is the highest for Scale 1 (23 items).

Table 4
Reliability Estimates for the IPE Components

Test Component	Reliability Indices		
	Split-Half Correlation[1]	Spearman-Brown Coefficient[2]	Coefficient Alpha[3]
Total IPE (51 items)	.769	.869	.829
Scale 1 (23 items)	.780	.877	.865
Scale 2 (11 items)	.546	.617	.518
Scale 3 (16 items)	.653	.790	.706

1 This internal consistency coefficient is based on the correlation between the odd and even items.
2 The Spearman-Brown coefficient is based on the assumption that the two halves of the IPE are strictly parallel.
3 The Coefficient Alpha is algebraically equivalent to Kuder-Richardson Formula 20 (KR20).

In order to understand the dimensionality of the ratings, an additional analysis was performed. Specifically, we first computed a correlation matrix over all raters for each scale. Next, we factor analyzed this matrix. This analysis was adapted from Sackett and Dreher (1982) who used a two step process to investigate the effectiveness of assessor judgments of exercise dimensions associated with assessment center methodology. In step 1, the dimension ratings for each of the 3 board raters were intercorrelated. The result was a multitrait-multimethod correlation matrix (MTMM) where the exercise score dimensions served as traits and methods were represented by raters (Campbell & Fiske, 1959). According to this methodology, the effectiveness of underlying dimensionality can be assessed by demonstrating both convergent and discriminant validity[4].

Step 2 required that the MTMM matrix be factor analyzed. The question in factor analytic terms was whether the factors represented raters or dimensions. Assuming both the orthogonal nature of the dimensions and the adequacy of the scales, our expectation was that the factors should identify checklist dimensions rather than rater behavior.

Using the minimum eigenvalue criterion (a factor needs a eigenvalue of 1 to be retained), a three factor solution was

---

[4] In this case, "reliability" rather than "convergent validity" should be used. Since the raters are using the same method (scales) to measure a single dimension, the analysis reflects agreement between two independent estimations using the same method (i.e., interrater reliability).

identified. Table 5 contains the rotated factor solution (i.e., varimax). As expected the factors represented checklist dimensions rather than rater dimensions.

Table 5
Rotated Factor Loadings for the Rater IPE Ratings

	Factor 1	Factor 2	Factor 3
Checklist 1			
Rater 1	**.972**	.112	.157
Rater 2	**.968**	.110	.168
Rater 3	**.972**	.110	.155
Checklist 3			
Rater 1	.100	**.978**	.118
Rater 2	.110	**.975**	.130
Rater 3	.118	**.973**	.127
Checklist 5			
Rater 1	.143	.135	**.950**
Rater 2	.168	.123	**.945**
Rater 3	.161	.115	**.947**

For example, Factor 1 consists of loadings associated with Checklist 1 (important aspects of the investigation). Factor 2 has the highest loadings related to Checklist 3. Factor 3 represents Checklist 5.

Discussion

This paper has attempted to describe a work sample methodology for selecting detectives for the Chicago Police Force. It began with an explanation of rationale for such an instrument. Next, we discussed the development and utilization of the IPE. We also discussed the reasons from which a cutscore was achieved. Finally, we examined the psychometric properties

of the test once it was used. However, there are several important issues that were not addressed.

The first relates to the exploratory nature of this procedure. We brought together a number of different methodologies in an attempt to build a model of a procedure that would effective in a relatively unstructured jobs. For example, we employed the work sample method while also borrowing from a structured, situational interview. Given the preliminary evidence, there seems to be some merit in this approach.

In addition, we would also caution that development costs associated with this procedure were high. Development time, the use of a number of detective, supervisors, and the final use of 50 oral boards was expensive. Cheaper and more efficient use of this type of procedure are possible. For example, the use of videotape cameras could dramatically reduce the cost of assessment. These tapes could be taken and reviewed by a small number of trained (or for that matter, expert) raters, who could examine ratee performance at their convenience..

We also did not discuss the results in terms of group fairness. It would be sufficient to say that there were significant mean group differences between majority and minority candidates. These differences are consistent with other cognitive ability tests. Given the nature of the procedure, these were expected.

The final and most important is that, to date, we were unable to develop criterion-related validity evidence. There

were several reasons for this. The first was that the test was used to promote only 32 officers to detective. This small sample size seriously restricts the power of any statistical test (Schmidt, Hunter & Ury, 1976). Second, the sufficiency of the only available criterion measure, a global trait oriented measure, was questionable. What must be remembered is that the nature of this test was more suited to determine if the underlying constructs (identified by the three scales) were related to detective performance. Hence, using a composite criterion may mask any construct validity evidence.

In summary, this test represents a step in the progression of work-sample, situational tests. And while there was a genuine feeling of the test's job-relatedness by raters and applicants, what needs to be done is to evaluate the relative utility of the procedure. The utility of the procedure is perhaps the optimal way of evaluating any type of selection procedure.

REFERENCES

Allen, M.J., & Yen, M.W. (1979). Introduction to Measurement Theory. Monterey, CA: Brooks/Cole Publishing Co.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105

Conley, P.R. (1987). The Youth Officer Examination: A Critique. Technical Report 87-901, Chicago Police Department.

Division of Industrial-Organizational Psychology, American Psychological Association (1985). Principles for the Validation and Use of Personnel Selection Procedures. Dayton OH: APA.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. Washington, DC: U.S. Government Printing Office.

Tukey, J.W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Sackett, P.R., & Dreher, G.F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. Journal of Applied Psychology, 67, 401-410.

Schmidt, F.L., Hunter, H.E., & Urry, V.E. (1976). Statistical power in criterion-related validation studies, Journal of Applied Psychology, 61, 473-485.